Occam factors and model independent Bayesian learning of continuous distributions

Ilya Nemenman[1,2] and William Bialek[2]

[1]*Department of Physics, Princeton University, Princeton, New Jersey 08544*
[2]*NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540*
(Received 11 September 2000; published 24 January 2002)

Learning of a smooth but nonparametric probability density can be regularized using methods of quantum field theory. We implement a field theoretic prior numerically, test its efficacy, and show that the data and the phase space factors arising from the integration over the model space determine the free parameter of the theory ("smoothness scale") self-consistently. This persists even for distributions that are atypical in the prior and is a step towards a model independent theory for learning continuous distributions. Finally, we point out that a wrong parametrization of a model family may sometimes be advantageous for small data sets.

I. INTRODUCTION

One of the central problems in learning is to balance "goodness of fit" criteria against the complexity of models. An important development in the Bayesian approach was thus the realization that there does not need to be any extra penalty for model complexity. If we compute the total probability that data are generated by a model, there is a factor from the volume in parameter space—the "Occam factor"— that discriminates against models with more parameters [1,2] or, more specifically, against models that are more complex in a precise information theoretic sense [3]. This works remarkably well for systems with a finite number of parameters and creates a complexity "razor" (after "Occam's razor") that is almost equivalent to the celebrated minimal description length (MDL) principle [4]. In addition, if the *a priori* distributions involved are strictly Gaussian, the ideas have also been proven to apply to some infinite-dimensional (nonparametric) problems [6]. It is not clear, however, what happens if we leave the finite-dimensional setting to consider nonparametric problems that are not Gaussian, such as the estimation of a smooth probability density. A possible route to progress on the nonparametric problem was opened by noticing [5] that a Bayesian prior for density estimation is equivalent to a quantum field theory (QFT). In particular, there are field theoretic methods for computing the infinite-dimensional analog of the Occam factor, at least asymptotically for large numbers of examples. These observations have led to a number of papers [7–10] exploring alternative formulations and their implications for the speed of learning. Here we return to the original formulation of Ref. [5] and address some of the questions left open by the previous work [11]. What is the result of balancing the infinite-dimensional Occam factor against the goodness of fit? Is the QFT inference optimal in using all of the information relevant for learning [3]? What happens if our learning problem is strongly atypical of the prior distribution?

The conclusions we finally make were not expected by us at the start of the project, and they will probably be not intuitively obvious to most of our readers either. Thus we chose to present this work in the same way it had originally proceeded. First, we develop a numerical scheme for implementation of the learning algorithm of Ref. [5]. Then we show some results of Monte Carlo simulations with this algorithm and notice some peculiar features that have not been predicted by the previous literature. Concurrently with the simulations, we present a simple analytical argument that explains these unexpected but extremely desirable features.

II. PRELIMINARIES

Following Ref. [5], if N independent, identically distributed samples $\{x_i\}, i=1,\ldots,N$, are observed, then the probability that a particular density $Q(x)$ gave rise to these data is given by

$$P[Q(x)|\{x_i\}] = \frac{\mathcal{P}[Q(x)]\prod_{i=1}^{N} Q(x_i)}{\int [dQ(x)]\mathcal{P}[Q(x)]\prod_{i=1}^{N} Q(x_i)}, \quad (1)$$

where $\mathcal{P}[Q(x)]$ encodes our *a priori* expectations of Q. Specifying this prior on a space of functions defines a QFT, and the optimal least square estimator is then the *a posteriori* Bayesian average

$$Q_{\text{est}}(x|\{x_i\}) = \frac{\langle Q(x)Q(x_1)Q(x_2)\cdots Q(x_N)\rangle^{(0)}}{\langle Q(x_1)Q(x_2)\cdots Q(x_N)\rangle^{(0)}}, \quad (2)$$

where $\langle\cdots\rangle^{(0)}$ means averaging with respect to the prior. Since $Q(x)\geq 0$, it is convenient to define an unconstrained field $\phi(x)$, $Q(x)\equiv(1/\ell_0)\exp[-\phi(x)]$, where the choice of the dimension setting constant ℓ_0 must not influence any final results. Other definitions are also possible [7], but we think that most of our results do not depend on this choice.

Next we should select a prior that regularizes the infinite number of degrees of freedom and allows learning. We want the prior $\mathcal{P}[\phi]$ to make sense as a continuous theory, independent of discretization of x on small scales. Since it is not clear what a renormalization procedure for a probability density would mean, we also require that when we estimate the distribution $Q(x)$ the answer must be everywhere finite. These conditions imply that our field theory must be ultraviolet (UV) convergent. For x in one dimension, a minimal choice is

$$\mathcal{P}[\phi(x)] = \frac{\exp\left[-\frac{\ell^{2\eta-1}}{2}\int dx\left(\frac{\partial^\eta \phi}{\partial x^\eta}\right)^2\right]}{\mathcal{Z}}$$

$$\times \delta\left[\frac{\int dx\, e^{-\phi(x)}}{\ell_0} - 1\right], \tag{3}$$

where $\eta > 1/2$, $\mathcal{Z}$ is the normalization constant, and the δ function enforces normalization of Q. We refer to ℓ and η as the *smoothness scale* and the *exponent*, respectively. They would be called *hyperparameters* in other machine learning literature [6].

In [5] this theory was solved for large N and $\eta = 1$ using the familiar WKB techniques. The saddle point (or the classical solution) for the ϕ averaging in $\langle \Pi_{i=1}^N Q(x_i)\rangle^{(0)}$ was found to be given by

$$\ell\partial_x^2\phi_{cl}(x) + \frac{N}{\ell_0}e^{-\phi_{cl}(x)} = \sum_{j=1}^N \delta(x-x_j), \tag{4}$$

and the fluctuation determinant around this saddle is

$$R = \exp\left[-\frac{1}{2}\sqrt{\frac{N}{\ell\ell_0}}\int dx\, e^{-\phi_{cl}(x)/2}\right]. \tag{5}$$

Then the correlation functions take a familiar form

$$\left\langle \prod_{i=1}^N Q(x_i)\right\rangle^{(0)} \approx \frac{1}{\ell_0^N}\exp(-S_{eff}[\phi_{cl}(x);\{x_i\}]), \tag{6}$$

$$S_{eff} = \frac{\ell}{2}\int dx(\partial_x\phi_{cl})^2 + \sum_{j=1}^N \phi_{cl}(x_j) - \ln R, \tag{7}$$

In Ref. [5] it was shown that, with such correlation functions, Eq. (2) is a "proper" solution to the learning problem. It is nonsingular even at finite N, it converges to the target distribution $P(x)$ that actually generates the data, and the variance of fluctuations around the target, $\psi(x) \equiv -\ln Q_{est}(x) - [-\ln \ell_0 P(x)]$, falls off rather quickly as $\sim 1/\sqrt{\ell NP(x)}$. It was also noted that the effective action [Eq. (7)] has acquired a term $-\ln R$, which grows as ℓ decreases. This is contrary to the data contribution, $\Sigma_{j=1}^N \phi_{cl}(x_j)$, which favors small ℓ and the corresponding overfitting. Thus the $-\ln R$ term may be rightfully called an infinite-dimensional generalization of the Occam factors. The authors speculated that, if the actual ℓ is unknown, one may average over it and hope that, much as in Bayesian model selection [1,2], the competition between the data and the fluctuations will select the optimal smoothness scale ℓ^*. Finally, they suggested that this optimal scale might behave as $\ell^* \sim N^{1/3}$.

Before we proceed to the numerical implementation of the above algorithm, a note is in order. At first glance the theory we study seems to look almost exactly similar to a Gaussian process [6]. This impression is produced by a Gaussian form of the smoothness penalty in Eq. (3), and by the fluctuation
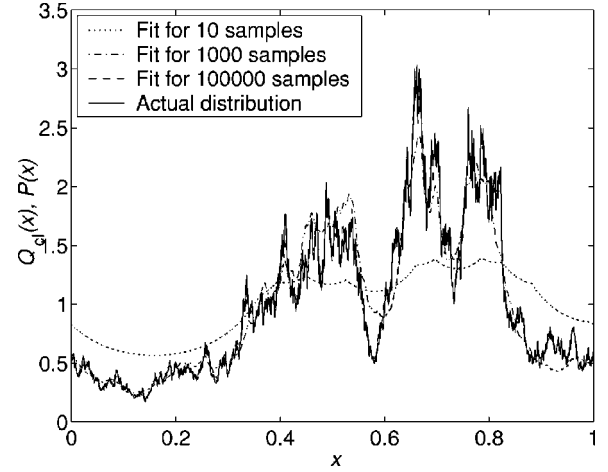

FIG. 1. Q_{cl} found for different N at $\ell = 0.2$.

determinant that plays against the goodness of fit in the smoothness scale (model) selection. However, both similarities are incomplete. The Gaussian penalty in the prior is amended by the normalization constraint, which gives rise to the exponential term in Eq. (4), and violates many familiar results that hold for Gaussian processes, the representer theorem [12] being just one of them. In the semiclassical limit of large N, Gaussianity is restored approximately, but the classical solution is extremely nontrivial, and the fluctuation determinant is only the leading term of the Occam's razor, not the complete razor as it is for a Gaussian process. In addition, it depends on the data only through the classical solution; this is remarkably different from the usual determinants arising in the Gaussian processes literature [6,7].

III. THE ALGORITHM

Numerical implementation of the theory is rather simple. First, to eliminate a possible infrared singularity in Eq. (5) [3,11], we confine x to a box $0 \le x \le L$ with periodic boundary conditions. The boundary value problem, Eq. (4), is then solved by a standard "relaxation" (or Newton) method of iterative improvements to a guessed solution [13] (for the target precision we always use 10^{-5}). The independent variable $x \in [0,1]$ is discretized in equal steps (10^4 for Figs. 1–4, and 10^5 for Figs. 5 and 6). We use an equally spaced grid to ensure stability of the method, while small step sizes are needed since the scale for variation of $\phi_{cl}(x)$ is [5]

$$\delta x \sim \sqrt{\ell/NP(x)}, \tag{8}$$

which can be rather small for large N or small ℓ.

Since the theory is UV convergent, we can generate random probability densities chosen from the prior Eq. (3) by replacing ϕ with its Fourier series and truncating the latter at some sufficiently high wave number k_c ($k_c = 1000$ for Figs. 1–4, and 5000 for Figs. 5 and 6). Then Eq. (3) enforces the amplitude of the kth mode ($k > 0$) to be distributed *a priori* normally around zero with the standard deviation

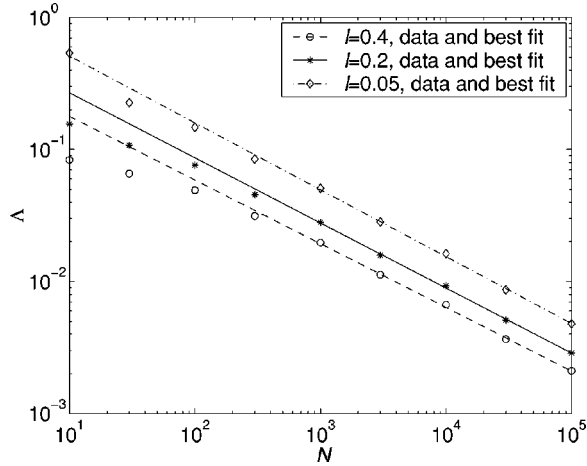$$\sigma_k = \frac{2^{1/2}}{\ell^{\eta-1/2}}\left(\frac{L}{2\pi k}\right)^\eta. \tag{9}$$

FIG. 2. Λ as a function of N and ℓ. The best fits are: for $\ell = 0.4$, $\Lambda = (0.54 \pm 0.07)N^{-0.483 \pm 0.014}$; for $\ell = 0.2$, $\Lambda = (0.83 \pm 0.08)N^{-0.493 \pm 0.09}$; for $\ell = 0.05$, $\Lambda = (1.64 \pm 0.16)N^{-0.507 \pm 0.09}$.

Once all these amplitudes are selected, the $k = 0$ harmonic is then set by the normalization condition.

Coded in such a way, the simulations are extremely computationally intensive because each iteration steps involves an inversion of a large matrix. Therefore, Monte Carlo averagings given here are only over 500 runs, fluctuation determinants are calculated according to Eq. (7), not using numerical path integration, and $Q_{\mathrm{cl}} = (1/\ell_0)\exp[-\phi_{\mathrm{cl}}]$ is always used as an approximation to Q_{est}.

IV. SIMULATIONS: CORRECT PRIOR

As an example of the algorithm's performance, Fig. 1 shows one particular learning run for $\eta = 1$ and $\ell = 0.2$. We see that singularities and overfitting are absent even for N as low as ten. Moreover, the approach of $Q_{\mathrm{cl}}(x)$ to the actual distribution $P(x)$ is remarkably fast: for $N = 10$, they are similar; for $N = 1000$, very close; for $N = 100\,000$, one needs to look carefully to see the difference between the two.
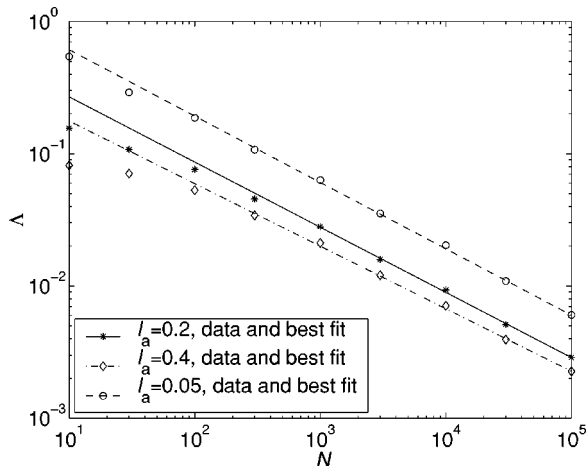
To quantify this similarity of distributions, we compute

FIG. 3. Λ as a function of N and ℓ_a. Best fits are: for $\ell_a = 0.4$, $\Lambda = (0.56 \pm 0.08)N^{-0.477 \pm 0.015}$; for $\ell_a = 0.05$, $\Lambda = (1.90 \pm 0.16)N^{-0.502 \pm 0.008}$. Learning is always with $\ell = 0.2$.
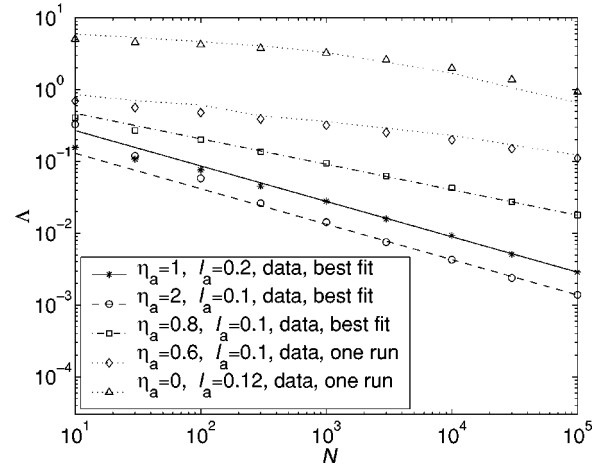

FIG. 4. Λ as a function of N, η_a, and ℓ_a. Best fits: for $\eta_a = 2$, $\ell_a = 0.1$, $\Lambda = (0.40 \pm 0.05)N^{-0.493 \pm 0.013}$; for $\eta_a = 0.8$, $\ell_a = 0.1$, $\Lambda = (1.06 \pm 0.08)N^{-0.355 \pm 0.008}$, $\ell = 0.2$ for all graphs, but the one with $\eta_a = 0$, for which $\ell = 0.1$.

the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(P||Q_{\mathrm{est}})$ between the true distribution $P(x)$ and its estimate $Q_{\mathrm{est}}(x)$, and then average over the realizations of the data points and the true distribution. As discussed in [3], this learning curve $\Lambda(N)$ measures the (average) excess cost incurred in coding the $N+1$th data point because of the finiteness of the data sample, and thus can be called the "universal learning curve." If the inference algorithm uses all of the information contained in the data that is relevant for learning ("predictive information" [3]), then [3,5,10,11]

$$\Lambda(N) \sim (L/\ell)^{1/2\eta}N^{1/2\eta - 1}. \tag{10}$$

We test this prediction against the learning curves in the actual simulations. For $\eta = 1$ and $\ell = 0.4, 0.2, 0.05$, these are shown in Fig. 2. One sees that the exponents are extremely close to the expected one-half, and the ratios of the prefactors are within the errors from the predicted scaling $\sim 1/\sqrt{\ell}$. All of this means that the proposed algorithm for finding
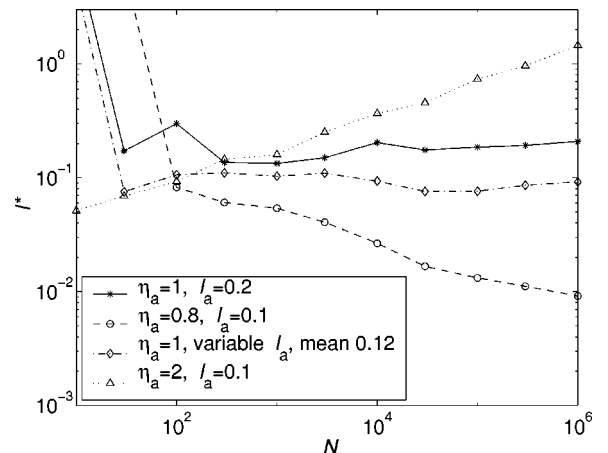

FIG. 5. Smoothness scale selection by the data. The lines that go off the axis for small N symbolize that S_{eff} monotonically decreases as $\ell \to \infty$.
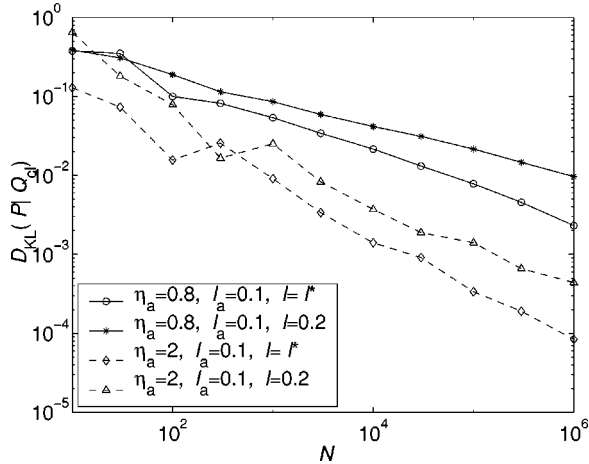
FIG. 6. Comparison of learning speed for the same data sets with different *a priori* assumptions.

densities not only works, but is at most a constant factor away from being optimal in using the predictive information of the sample set.

V. SIMULATIONS: WRONG PRIOR

Next we investigate how one's choice of the prior influences learning. We first stress that there is no such thing as a *wrong prior*. If one admits a possibility of it being wrong, then it does not encode all of the *a priori* knowledge. It does make sense, however, to ask what happens if the distribution we are trying to learn is an extreme outlier in the prior $\mathcal{P}[\phi]$. One way to generate such an example is to choose a typical function from a different prior $\mathcal{P}'[\phi]$, and this is what we mean by "learning with a wrong prior."

If the prior is wrong in this sense, and learning is described by Eqs. (2)–(4), then we still expect the asymptotic behavior, Eq. (10), to hold; only the prefactors of Λ should change, and those must increase since there is an obvious advantage in having the right prior; we illustrate this in Figs. 3 and 4.

For Fig. 3, both $\mathcal{P}'[\phi]$ and $\mathcal{P}[\phi]$ are given by Eq. (3), but $\mathcal{P}'$ has the "actual" smoothness scale $\ell_a = 0.4, 0.05$, and for $\mathcal{P}$ the "learning" smoothness scale is $\ell = 0.2$ (we show the case $\ell_a = \ell = 0.2$ again as a reference). The $\Lambda \sim 1/\sqrt{N}$ behavior is seen unmistakably. The prefactors are a bit larger (unfortunately, insignificantly) than the corresponding ones from Fig. 2, so we may expect that the "right" ℓ, indeed, provides better learning (see later for a detailed discussion).

Further, Fig. 4 illustrates learning when not only *l*, but also η is "wrong" in the sense defined above. We illustrate this for $\eta_a = 2, 0.8, 0.6, 0$ (remember that only $\eta_a > 0.5$ removes UV divergences). Again, the inverse square root decay of Λ should be observed, and this is evident for $\eta_a = 2$. The $\eta_a = 0.8, 0.6, 0$ cases are different. Even for *N* as high as 10^5 the estimate of the distribution is far from the target, thus the asymptotic regime is not reached. This is a crucial observation for our subsequent analysis of the smoothness scale determination from the data. Remarkably, Λ (both averaged and in the single runs shown) is monotonic, so even in the cases of *qualitatively* less smooth distributions *there still is*

no overfitting. On the other hand, Λ is well above the asymptote for $\eta = 2$ and small *N*, which means that initially too many details are expected and wrongfully introduced into the estimate, but then they are almost immediately ($N \sim 300$) eliminated by the data.

VI. SMOOTHNESS SCALE SELECTION

Following the argument suggested in [5], we now view $\mathcal{P}[\phi]$, Eq. (3), as being a part of some wider model that involves a prior over ℓ. The details of the prior are irrelevant, however, if $S_{\text{eff}}(\ell)$, Eq. (7), has a minimum that steepens as *N* grows. We explicitly note that this mechanism is *not* a tuning of the prior's parameters, but Bayesian inference at work: ℓ^* emerges in a competition between the kinetic, the data, and the Occam terms to make S_{eff} smaller, and thus the *total* probability of the data is larger. In its turn, larger probability means, roughly speaking, a shorter total code length, hence, the relation to the MDL paradigm [4].

The data term, on average, is equal to $ND_{\text{KL}}(P||Q_{\text{cl}})$, and, for very regular $P(x)$ (an implicit assumption in [5]), it is small. Thus only the kinetic and the Occam terms matter, and $\ell^* \sim N^{1/3}$ [5]. For less regular distributions $P(x)$, this is not true [cf. Fig. (4)]. For $\eta = 1$, $Q_{\text{cl}}(x)$ approximates large-scale features of $P(x)$ very well, but details at scales smaller than $\sim \sqrt{\ell/NL}$ are averaged out. If $P(x)$ is taken from the prior, Eq. (3), with some η_a, then these details fall off with the wave number *k* as $\sim k^{-\eta_a}$. Thus the data term is $\sim N^{1.5 - \eta_a} \ell^{\eta_a - 0.5}$ and is not necessarily small. For $\eta_a < 1.5$ this dominates the kinetic term and competes with the fluctuations to set

$$\ell^* \sim N^{(\eta_a - 1)/\eta_a}, \qquad \eta_a < 1.5. \qquad (11)$$

There are two remarkable things about Eq. (11). First, for $\eta_a = 1$, ℓ^* stabilizes at some constant value, which we expect to be equal to ℓ_a. Second, even for $\eta \neq \eta_a$, Eqs. (10) and (11) ensure that Λ scales as $\sim N^{1/2\eta_a - 1}$, which is at worst a constant factor away from the best scaling, Eq. (10), achievable with the "right" prior, $\eta = \eta_a$. So, by allowing ℓ^* to vary with *N* we can correctly capture the structure of models that are qualitatively different from our expectations ($\eta \neq \eta_a$) and produce estimates of *Q* that are extremely robust to the choice of the prior. To our knowledge, this feature has not been noted before in a reference to a nonparametric problem.

We present simulations relevant to these predictions in Figs. 5 and 6. Unlike in the previous figures, the results are not averaged due to extreme computational costs, so all our further claims have to be taken cautiously. On the other hand, selecting ℓ^* in single runs has some practical advantages, we are able to ensure the best possible learning for any realization of the data. Figure 5 shows single learning runs for various η_a and ℓ_a. In addition, to keep the figure readable, we do not show runs with $\eta_a = 0.6, 0.7, 1.2, 1.5, 3$, and $\eta_a \rightarrow \infty$, which is a finitely parameterizable distribution. All of these display a good agreement with the predicted scalings, Eq. (11) for $\eta_a < 1.5$, and $\ell^* \sim N^{1/3}$ otherwise. Next we calculate the KL divergence between the target and the estimate

at $\ell = \ell^*$. The average of this divergence over the samples and the prior is the learning curve [cf. Eq. (10)]. For $\eta_a = 0.8, 2$ we plot the divergences in Fig. 6 side by side with their fixed $\ell = 0.2$ analogues. Again, the predictions clearly are fulfilled. Note, that for $\eta_a \neq \eta$ there is a *qualitative* advantage in using the data induced smoothness scale.

VII. PARAMETRIZATION AS A WRONG PRIOR

The last four figures have illustrated some aspects of learning with ''wrong'' priors. However, all of our results may be considered as belonging to the ''wrong prior'' class. Indeed, the actual probability distributions we used were not nonparametric continuous functions with smoothness constraints, but were composed of k_c Fourier modes, thus had $2k_c$ parameters. For finite parameterization, asymptotic properties of learning usually do not depend on the priors (cf. [3,4]), and priorless theories can be considered [14]. In such theories it would take well over $2k_c$ samples to even start to close down on the actual value of the parameters, and yet a lot more to get accurate results. However, using the wrong continuous parameterization $[\phi(x)]$ we were able to obtain good fits for as low as 1000 samples (cf. Fig. 1) with the help of the prior Eq. (3). Moreover, learning happened continuously and monotonically without huge chaotic jumps of overfitting that necessarily accompany any brute force parameter estimation method at low N. So, for some cases, a *seemingly more complex model* is actually *easier* to learn.

Thus our claim: when data are scarce and the parameters are abundant, one gains even by using the regularizing powers of wrong priors. The priors select some large scale features that are the most important to learn first and fill in the details as more data become available (see [3] on relation of this to the Structural Risk Minimization theory). If the global features are dominant (arguably, this is generic), one actually wins in the learning speed (cf. Figs. 2, 3, and 6). If, however, small scale details are as important, then one at least is guaranteed to avoid overfitting (cf. Fig. 4).

One can summarize this in an Occam-like fashion [3]. If two models provide equally good fits to data, *a simpler one should always be used*. In particular, the predictive information, which quantifies complexity [3], and of which Λ is the derivative, in a QFT model is $\sim N^{1/2 \eta}$, and it is $\sim k_c \ln N$ in the parametric case. So, for $k_c > N^{1/2 \eta}$, one should prefer a ''wrong'' QFT formulation to the correct one. These results are very much in the spirit of our whole program. Not only is the value of ℓ^* selected that simplifies the description of the data, but the continuous parameterization itself serves the same purpose.

VIII. SUMMARY

The field theoretic approach to density estimation not only regularizes the learning process but also allows the self-consistent selection of smoothness criteria through an infinite-dimensional version of the Occam factors. We have shown numerically, and then explained analytically that this works, even more clearly than was conjectured. For $\eta_a < 1.5$, Λ truly becomes a property of the data, and not of the Bayesian prior. If we can extend these results to other η_a and combine this work with the reparametrization invariant formulation of [8,9], this should give a complete theory of Bayesian learning for one dimensional distributions, and this theory has no arbitrary parameters. In addition, if this theory properly treats the limit $\eta_a \rightarrow \infty$, we should be able to see how the well-studied finite-dimensional Occam factors and the MDL principle arise from a more general nonparametric formulation.

[1] D. MacKay, Neural Comput. **4**, 415 (1992).

[2] V. Balasubramanian, Neural Comput. **9**, 349 (1997); e-print cond-mat/9601030.

[3] W. Bialek, I. Nemenman, and N. Tishby, Neural Comput. **13**, 2409 (2001); e-print physics/0007070.

[4] J. Rissanen, *Stochastic Complexity and Statistical Inquiry* (World Scientific, Singapore, 1989).

[5] W. Bialek, C. Callan, and S. Strong, Phys. Rev. Lett. **77**, 4693 (1996); e-print cond-mat/9607180.

[6] D. MacKay, *Advances in Neural Information Processing Systems 10*, Tutorial Lecture Notes (unpublished). ftp:// wol.ra.phy.cam.ac.uk/pub/mackay/gp.ps.gz .

[7] T. Holy, Phys. Rev. Lett. **79**, 3545 (1997); e-print physics/9706015.

[8] V. Periwal, Phys. Rev. Lett. **78**, 4671 (1997); e-print hep-th/9703135.

[9] V. Periwal, Nucl. Phys. B **554**, 719 (1999); e-print adap-org/9801001.

[10] T. Aida, Phys. Rev. Lett. **83**, 3554 (1999); e-print cond-mat/9911474.

[11] A more detailed version of our current analysis may be found in I. Nemenman, Ph.D. thesis, Princeton University, Princeton, 2000; e-print physics/0009032.

[12] G. Wahba, *Advances in Kernel Methods—Support Vector Learning*, edited by B. Schölkopf, C. J. S. Burgess, and A. J. Smola (MIT Press, Cambridge, MA, 1999), pp. 69–88; ftp:// ftp.stat.wisc.edu/pub/wahba/nips97rr.ps .

[13] W. Press *et al.*, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1988).

[14] V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998).

[15] I. Nemenman and W. Bialek, *Advances in Neural Information Processing Systems 13*, edited by T. K. Leen, T. G. Dietterich, and V. Tresp (MIT Press, Cambridge, MA, 2001), pp. 75–81; e-print cond-mat/0009165.